MEASURING FORECAST SKILL:
IS IT REAL SKILL OR
IS IT THE VARYING CLIMATOLOGY?

Thomas M. Hamill*

NOAA Earth System Research Laboratory, Physical Sciences Division

Boulder, Colorado

Josip Juras

Geophysical Institute, Faculty of Science, University of Zagreb

Zagreb, Croatia

Submitted to *Quarterly Journal of the Royal Meteorological Society*

Revised

5 May 2006

*Corresponding author address:

Dr. Thomas M. Hamill
NOAA Earth System Research Lab, Physical Sciences Division
R/PSD 1, 325 Broadway
Boulder, CO 80305-3328 USA

e-mail: tom.hamill@noaa.gov
phone: 001 (303) 497-3060

SUMMARY

It is common practice to summarize the skill of weather forecasts from an accumulation of samples spanning many locations and dates. In calculating many of these scores, there is an implicit assumption that the climatological frequency of event occurrence is approximately invariant over all samples. If the event frequency actually varies among the samples, the metrics may report a skill that is different than expected. Many common deterministic verification metrics such as threat scores are prone to misreporting skill, and probabilistic forecast metrics such as the Brier skill score and relative operating characteristic skill score can also be affected.

Three examples are provided that demonstrate unexpected skill, two from synthetic data and one with actual forecast data. In the first example, positive skill was reported in a situation where metrics were calculated from a composite of forecasts that were comprised of random draws from the climatology of two distinct locations. As the difference in climatological event frequency between the two locations was increased, the reported skill also increased. A second example demonstrates that when the climatological event frequency varies among samples, the metrics may excessively weight samples with the greatest observational uncertainty. A final example demonstrates unexpectedly large skill in the equitable threat score of deterministic precipitation forecasts.

Guidelines are suggested for how to adjust skill computations to minimize these effects.

Keywords: forecast verification, ensemble forecasting, probabilistic weather forecasts, equitable threat score, relative operating characteristic, Brier skill score, contingency tables

1. **Introduction**

This article will demonstrate that many commonly used weather forecast verification metrics are capable of reporting positive forecast skill in situations where the meteorologist would assume none truly exists, or the metrics may report different skill than expected. Depending on the metric and the situation, this effect can be large or small. The unexpected skill is a consequence of inappropriately pooling data over subsets with different climatological event frequencies.

Our interest in this topic resulted from using conventional verification metrics and diagnosing unexpectedly large skill. For example, the first author used a common probabilistic metric, the relative operating characteristic, in a comparison of ensemble forecast methods (Hamill et al. 2000, Fig. 13). The author reported a relative operating characteristic curve for wind speed forecasts at 5 days lead that indicated a highly skillful forecast, different than experience would suggest for this lead time. The second author discussed the unexpectedly large forecast skill (Juras 2000) in a comment on a Buizza et al. (1999) article. It was indicated that the chosen metrics might report unexpectedly large skill if climatological event frequencies varied within the verification area. This issue was also raised in Mason (1989) and less directly in other meteorological publications, including Buizza (2001; p. 2335), Stefanova and Krishnamurti (2002, p. 543), Atger (2003), Glahn (2004; p. 770), and Göber et al. (2004). Still, there are many published articles that may have applied common verification metrics, incorrectly assuming that the conventional method of calculation would result in zero skill for the reference, which is commonly assumed to be a random draw from the observed climatological distribution.

Here, section 2 will provide a brief review of the three chosen metrics that may be subject to misestimating skill, the Brier skill score (Wilks 1995, p. 260), the relative operating characteristic (Swets 1973, Harvey et al. 1992) skill score, and the equitable threat score (Schaefer 1990). Many other metrics such as the ranked probability skill score (Epstein 1969, Murphy 1971, Wilks 1995, p. 272), economic value diagrams (Richardson 2000, Palmer et al. 2000, Richardson 2001b, Zhu et al. 2002, and Buizza et al. 2003), and other contingency-table based threat scores will not be discussed but can be assumed to be subject to the same effect. In addition to describing the conventional method of calculation of these metrics, Section 2 will also describe possible improved methods of calculation. Section 3 follows with two simple examples of how unexpected skill can be diagnosed from synthetic weather data when using the conventional methods of calculation. Section 4 demonstrates how large the mis-estimation effect can be for a common real-weather verification problem, the threat scores of short-range precipitation forecasts. Section 5 concludes with a discussion of the implications and how to adapt verification strategies to minimize or avoid this effect.

2. **Review of three common verification metrics**

Below, three general verification metrics are reviewed, the Brier skill score, relative operating characteristic (*ROC*) skill score, and the equitable threat score.

The long-used Brier score (Brier 1950) is a measure of the mean-square error of probability forecasts for a dichotomous (two-category) event, such as the occurrence/non-occurrence of precipitation. A review is provided in Wilks (2006, p. 284), and references therein provide further background. The Brier score is often converted to a skill score, its value normalized by the Brier score of a reference forecast such as climatology (ibid). A

Brier skill score (*BSS*) of 1.0 indicates a perfect probability forecast, while a BSS of 0.0 should indicate the skill of the reference forecast (see Mason 2004 for further discussion of whether a *BSS* of 0.0 indicates no skill).

The relative operating characteristic (*ROC*) has gained widespread acceptance in the past few years as a tool for probabilistic weather forecast verification. The *ROC* has been used for decades in engineering, biomedical, and psychological applications. The *ROC* measures the hit rate of a forecast against its false-alarm rate as the decision threshold (perhaps a quantile of a probabilistic forecast) is varied. It also can be understood as a graph of the tradeoff of Type I vs. Type II statistical errors in a hypothesis test (Swets 1973). The *ROC's* application in meteorology was proposed in Mason (1982), Stanski et al. (1989), and Harvey et al. (1992). The *ROC* was recently made part of the World Meteorological Organization's (WMO) standard (WMO, 1992). Characteristics of the *ROC* have been discussed in Buizza et al. (1998), Mason and Graham (1999, 2002), Juras (2000), Wilson (2000), Buizza et al. (2000ab), Wilks (2001), Kheshgi and White (2001), Kharin and Zwiers (2003), Mason (2003), and Marzban (2004). The technique has been used to diagnose ensemble forecast accuracy in, for example, Buizza and Palmer (1998), Buizza et al. (1999), Hamill et al. (2000), Palmer et al. (2000), Richardson (2000, 2001ab), Wandishin et al. (2001), Ebert (2001), Mullen and Buizza (2001, 2002), Bright and Mullen (2002), Yang and Arritt (2002), Legg and Mylne (2004), Zhu et al. (2002), Toth et al. (2003), and Gallus and Segal (2004). Harvey et al. (1992) provide a thorough review of the concepts underlying the *ROC*. In subsequent discussion, we will discuss the skill score "*ROCSS*" derived from the *ROC*.

The equitable threat score (*ETS*) provides one of many ways of summarizing the ability of a deterministic forecast to correctly forecast a dichotomous (two-category) event. The *ETS* will produce a score of 1.0 for a perfect forecast, and random forecasts should be assigned a value of 0.0. The *ETS* is commonly used to evaluate the skill of forecasts, especially precipitation. See, for example, Rogers et al. (1995, 1996), Hamill (1999), Bayler et al. (2000), Stensrud et al. (2000), Xu et al. (2001), Ebert (2001), Gallus and Segal (2001), Chien et al. (2002), and Accadia et al. (2003).

The methods for computing these metrics are now discussed, starting with the probabilistic metrics. The *BSS* and *ROC* will be generated from ensemble forecasts, though they can be generated from any probabilistic forecast.

Start by defining a dichotomous event of interest, such as occurrence/non-occurrence of precipitation, or temperature above or below a threshold. Let $\mathbf{X}_e(j) = [X_1(j), \ldots, X_n(j)]$ be an *n*-member ensemble forecast of the relevant scalar variable (again, precipitation or temperature) for the *j*th of *m* samples (taken over many case days and/or locations). The ensemble at that day and location is first sorted from lowest to highest. This sorted ensemble is then converted into an *n*-member binary forecast $\mathbf{I}_e(j) = [I_1(j), \ldots , I_n(j)]$ indicating whether the event was forecast (= 1) or not forecast (= 0) by each member. The observed weather is also converted to binary, denoted by $I_o(j)$.

(a) Brier skill scores

Assuming that each member forecast is equally likely, a forecast probability $p_f(j)$ for the *j*th sample is calculated from the binary ensemble forecasts:

$$p_f(j) = \frac{1}{n}\sum_{i=1}^{n} I_i(j) \quad . \tag{1}$$

The Brier score of the forecast BS_f is calculated as

$$BS_f = \frac{1}{m}\sum_{j=1}^{m}\left\{p_f(j) - I_o(j)\right\}^2 \quad . \tag{2}$$

A Brier skill score (*BSS*) is commonly calculated as

$$BSS = 1 - \frac{BS_f}{BS_c} \quad , \tag{3}$$

where BS_c is the Brier score of the reference probability forecast, commonly the probability of event occurrence from climatology.

Ideally, the climatological probabilities would be determined from independent data, but commonly they are calculated from the sample observed data. In the conventional method of calculation, an average climatology p_c is used:

$$p_c = \frac{1}{m}\sum_{j=1}^{m} I_o(j) \ , \tag{4}$$

in which case the reference Brier score of climatology used in eq. (3) is

$$BS_c = \frac{1}{m}\sum_{j=1}^{m}\left\{p_c - I_o(j)\right\}^2 \quad . \tag{5}$$

The conventional method of calculation of the BSS in eqs. (1) – (5) may report a score that differs from what the meteorologist may expect if the climatological event frequency is known to vary among the *m* samples (section 3). Consequently, we propose some alternative methods of formulation of the scores, and later we will discuss the change in skill that was reported under the new calculations.

Suppose the samples could be split up into n_c subsets, each with a distinct climatological event frequency. Let $p_c(k)$ be the climatological event frequency in the *k*th of the n_c subsets. Also, let there be $n_s(k)$ samples in this subset, and let $\mathbf{r}_k = [r(1)$, … ,

$r(n_s(k))$] be the associated set of sample indices from the *m* samples. Then suppose the Brier score of climatology is calculated separately for each subset with a different climatology:

$$\widetilde{BS}_c\left(k\right) = \frac{1}{n_s(k)} \sum_{j=1}^{n_s(k)} \left\{ p_c\left(k\right) - I_o\left(r(j)\right) \right\}^2 \quad . \tag{6}$$

A possible alternative calculation of the Brier score of climatology would then be to calculate a sample weighted average:

$$\widetilde{BS}_c = \sum_{k=1}^{n_c} \frac{n_s(k)}{m} \widetilde{BS}_c\left(k\right). \tag{7}$$

The *BSS* would then calculated following eq. (3), replacing BS_c with $\widetilde{BS}_c$.

A third possible alternative for calculating the *BSS* would be to calculate the Brier Score of the forecasts separately for the different subsets with climatological event frequencies, just as was done with the climatological forecast in eq. (6):

$$\widetilde{BS}_f\left(k\right) = \frac{1}{n_s(k)} \sum_{j=1}^{n_s(k)} \left\{ p_f\left(r(j)\right) - I_o\left(r(j)\right) \right\}^2 . \tag{8}$$

Then the *BSS* would be computed as a sample-weighted average of the skill scores for each distinct climatological regime:

$$\overline{BSS} = \sum_{k=1}^{n_c} \frac{n_s\left(k\right)}{m} \left(1 - \frac{\widetilde{BS}_f\left(k\right)}{\widetilde{BS}_c\left(k\right)} \right). \tag{9}$$

This may better conform with forecaster intuition, e.g., if two locations with equal numbers of samples have *BSS*es of 0.0 and 1.0, a skill of 0.5 will be reported.

(b) ROC diagrams and the ROC skill score

For ensembles, the *ROC* is a curve that indicates the relationship between hit rate and false alarm rate as different sorted ensemble members are used as decision thresholds. The area under the *ROC* curve can be used in the calculation of a probabilistic skill score. The conventional method of calculation of the *ROC* from ensembles typically starts with the population of 2x2 contingency tables, with separate contingency tables tallied for each sorted ensemble member. The contingency table has four elements: $\Gamma_i = [\ a_i, b_i, c_i, d_i]$, indicating the fraction of hits, false alarms, misses, and correct rejections when the value of the ith sorted member is used as the forecast (Table 1). The contingency table is populated using data over all *m* samples.

The hit rate (*HR*) for the *i*th sorted member forecast is defined as

$$HR_i = \frac{a_i}{a_i + c_i}. \tag{10}$$

Similarly, the false alarm rate is defined as

$$FAR_i = \frac{b_i}{b_i + d_i}. \tag{11}$$

This prototypical *ROC* is a plot of HR_i (ordinate) vs. FAR_i (abscissa), $i = 1, \ldots, n$. A *ROC* curve that lies along the diagonal *HR*=*FAR* line is commonly believed to indicate no skill; a curve that sweeps out maximal area, as far toward the upper left corner as possible, is believed to indicate maximal skill. The *ROC* is commonly summarized through the integrated area under the *ROC* curve, or *AUC*. A perfect forecast AUC_{perf} =1.0, and forecasts that are random draws from climatology are presumed to provide an

AUC_{clim} = 0.5. In order to calculate the forecast area AUC_f, for the *n*-member ensemble let us assume the existence of fictitious zeroth and $n+1$th ensemble members to provide boundary conditions HR_0 = 0.0, FAR_0 = 0.0, HR_{n+1} = 1.0, and FAR_{n+1} = 1.0. Then an approximate integral AUC_f can be calculated as

$$AUC_f = \sum_{i=1}^{n+1} \frac{\left(FAR_i - FAR_{i-1}\right)\left(HR_i + HR_{i-1}\right)}{2} \tag{12}$$

(there are other valid methods of calculation). Commonly, a skill score *ROCSS* is calculated from AUC_f (Wilks 2006, p. 295):

$$ROCSS = \frac{AUC_f - AUC_{c\mathrm{lim}}}{AUC_{perf} - AUC_{c\mathrm{lim}}} = \frac{AUC_f - 0.5}{1.0 - 0.5} = 2\,AUC_f - 1\ . \tag{13}$$

As will be demonstrated in section 3, the conventional method of calculation of the *ROC* and *ROCSS* can result in an estimation of skill where none was expected if the climatological event frequency varies among samples. Hence, we outline a possible alternative method of calculation of *ROC* area and skill. Assume as with the *BSS* that we can divide up the samples into n_c subsets with distinct climatological event frequencies. Then an alternative method for calculation of the *ROC* area would be to calculate it separately for each subgroup and produce a weighted-average *ROC* area, which we will call $\overline{AUC}_f$. Using the $n_s(k)$ samples in the *kth* subset, the hit rates and false alarm rates for the *kth* climatology are

$$\widetilde{HR}_i\left(k\right) = \frac{a_i\left(k\right)}{a_i\left(k\right) + c_i\left(k\right)} \tag{14}$$

and

$$\widetilde{FAR}_i\left(k\right) = \frac{b_i\left(k\right)}{b_i\left(k\right) + d_i\left(k\right)}. \tag{15}$$

From this, the area under the *ROC* curve for the *k*th subset can be calculated in a manner analogous to eq. (12), providing $\widetilde{AUC}_f(k)$. Then, as was done with the *BSS*, a sample-weighted $\overline{AUC}_f$ is calculated according to

$$\overline{AUC_f} = \sum_{k=1}^{n_c} \frac{n_s\left(k\right)}{m}\ \widetilde{AUC_f}\left(k\right)\ , \tag{16}$$

and a skill score is calculated using eq. (13), substituting $\overline{AUC}_f$ for AUC_f.

(c) Equitable threat score

Assume now that we are evaluating deterministic forecasts rather than ensembles. The conventional method of calculating the *ETS* assumes Table 1 is populated with all the samples available (here we drop the $_i$ subscript in Table 1 denoting the ensemble member number). The equation for the *ETS* is

$$ETS = \frac{a - a_r}{a + b + c - a_r}, \tag{17}$$

where a_r is the expected fraction of hits for a random forecast

$$a_r = \frac{\{a+c\}\,\{a+b\}}{a+b+c+d}. \tag{18}$$

As with the other scores, we shall show in sections 3 and 4 that this conventional method of calculation will produce an unexpectedly high estimate in situations where the climatology varies. An alternative method of calculation of the *ETS* respects the possibility of different regions with different climates. Again, assume we have n_c contingency tables, each associated with samples with a distinct climatological event frequency. For the *kth* distinct climatology we thus construct a separate contingency table and calculate the threat score $\widetilde{ETS}(k)$. An alternative, sample-weighted *ETS* is then calculated as

$$\overline{ETS} = \sum_{k=1}^{n_c} \frac{n_s(k)}{m} \widetilde{ETS}(k) \quad . \tag{19}$$

3. **Example of skill overestimation: synthetic data at two independent locations**

Using synthetic data, we now illustrate two general problems with verification metrics calculated in the conventional manner. First, they may report skill even when the forecasts are samples from the reference, no-skill climatology. This may occur when the overall sample is comprised of subsamples that are drawn from different climatological distributions. Second, when the climatological uncertainty of event occurrence varies among samples, the skill scores may reflect an uneven weighting of the sample data.

(a) Positive skill diagnosed from reference climatological forecasts.

Suppose a hypothetical planet was covered by a global ocean interrupted only by two small, isolated islands, and suppose island weather forecasting was utterly impossible on this planet; the best one can do was to forecast the climatological probability distribution appropriate to each island, which is assumed stationary. Given that the weather appears random to residents on each island, one would expect that a skill score should report zero skill, a desired attribute that is part of the property known as "equitability" (Gandin and Murphy 1992, Wilks 2006 p. 274).

To simulate this scenario, assume that at island 1, on each day the observed daily maximum temperature was a draw from a normal distribution with a mean of α and a standard deviation of 1.0, which we denote notationally as ~ $N(+\alpha, 1)$. We also generated a 100-member ensemble each day to calculate the *BSS* and *ROCSS* and a single-member deterministic forecast to calculate the *ETS*. In each instance the forecasts were also ~ $N(+\alpha, 1)$ and were uncorrelated with each other and with the observation. On island 2, each day's observed daily maximum temperature ~ $N(-\alpha, 1)$, and again a 100-member ensemble and deterministic forecast were drawn from ~ $N(-\alpha, 1)$, with uncorrelated forecasts and observations. We will consider the event that the temperature was greater than zero. Forty-thousand days of forecasts and observations were generated for island and each value of α, and we examine the skill scores as α increases from zero, that is, as the two islands' climatologies grow increasingly different.

Figure 1 synthesizes the forecasts scores' overestimate as a function of α when the Brier skill score was calculated by computing the pooled samples by eqs. (4) – (5), the *ROCSS* was calculated using eqs. (10) – (13), and the *ETS* was calculated using eqs. (17) – (18). Hereafter, these will be called "the conventional methods" of calculation. The expected skill should be zero regardless of the value of α , for forecasts were always drawn from the reference climatological distribution appropriate to that island. However, as α was increased, the diagnosed forecast skill increased as well.

What was the source of the skill estimates being larger than expected? On island 1, the climatological probability of the observed being greater than zero increased with α, while on island 2 it decreased. The probabilities estimated from the ensemble behaved similarly, increasing with α on island 1 and decreasing on island 2. However, each of the conventional methods of calculating the scores implicitly assumed that the reference climatological event probability *for all forecasts at all α was a fixed 0.5*, a consequence of pooling the data from both islands together. Hence as α increased, the randomly drawn forecasts became increasingly sharp and accurate relative to this nonspecific composite climatology. The random forecasts from each island were awarded higher and higher scores based merely on the increasing differences in the two islands' mean temperatures, not through any intrinsic improvement in forecast skill. This illustrates that these scores may report unexpectedly large skill in situations where the climatologies differ among the samples used to populate the contingency tables; they credit a forecast with having skill when the climatologies of the individual samples are different from the climatology of the combined samples. In this example, the more the climatologies differed, the larger the diagnosed skill.

Though not shown here, an overestimation of skill would still have occurred even if forecasts on each island were positively correlated with the observed value and thus skillful. In such a situation, the actual skill would have been inflated by an additional amount due to the compositing of the two islands' climatology. This inflation of skill also makes it more difficult to evaluate potential forecast improvements. When α was very large, a forecast was scored as nearly perfect regardless of whether or not the forecast actually was nearly perfect. The difference between good and mediocre forecasts is thus shrunk, complicating the task of evaluating whether one model was better than another.

Consequently, the preferred course of action when the underlying climatology varies among samples is to *analyze the data separately for each distinct climatological regime*. A similar and more general conclusion was arrived at in the classic paper on "Simpson's Paradox" (Simpson 1951; see comment 7 on second-order interactions[1]). Cochran (1954) also is unambiguous with regards to inferences from contingency tables:

> "*One method that is sometimes used is to combine all the data into a single 2x2 table … this procedure is legitimate only if the probability p of an occurrence (on the null hypothesis) can be assumed to be the same in all the individual 2x2 tables. Consequently, if p obviously varies from table to table, or we suspect that it may vary, this procedure should not be used.*"

Cochran also proposed a statistical test to determine if contingency table data can be added; Mantel and Haenszel (1959) proposed a related test, and Agresti (2002, p. 231) provided a summary. Unfortunately, the Cochran and Mantel-Haenszel tests may be difficult to apply in meteorological verification, for one of the underlying assumptions is that the samples used to populate the contingency tables are independent. In meteorological verification, two samples may come from adjacent grid points that will in fact have correlated errors.

The meteorological statistician may sometimes still desire a single-number summary of the skill of the forecast, especially if the sample size of the forecasts is

[1] Simpson actually asserts something even more rigorous; contingency-table data can be added only when there are no "second-order interactions" in the contingency tables. These interactions may occur due to differences in climatological event frequency, but they also may occur in situations where the forecast skills were different between the subsets.

limited in each region with a different climatological event frequency. To preserve the desirable property of ensuring that random draws from the no-skill reference are evaluated as having null skill, the method of calculating the skill scores could be reformulated or the problem could be transformed to eliminate the effect of the varying climatology. For example, had the *BSS* been calculated with eqs. (3), (6), and (7) or eqs. (6), (8), and (9), had the *ROCSS* been calculated with eqs. (13) – (16), and had the *ETS* been calculated separately at each island and then averaged using eq. (19), the reported scores would have been zero within sampling error. Another way to report the expected zero skill would be to change the test threshold to one where the climatological event frequencies were identical among sub-samples. For example, change the test threshold from "temperature greater than zero" to "exceeding the 50th percentile of each individual island's climatological distribution." Of course, reformulating the verification problem in this manner may not address the underlying question asked by the researcher.

(b) Skill contributions weighted toward samples with larger observed uncertainty.

This experimental setup will illuminate how samples with different underlying climatological uncertainty[2] can be unequally weighted, affecting the computation of skill. Our two-island scenario is now altered; consider the event that the daily maximum temperature was greater than 2.0. Island 1's observed maximum temperature was randomly drawn from a ~$N(0,1)$ distribution, the forecasts were also ~ $N(0,1)$, and

[2] Uncertainty here refers to a measure of the intrinsic variability of the observations, as in the Brier score decomposition (Wilks 2006, p. 286). Given a climatological event probability p_c, the uncertainty is $p_c(1-p_c)$, which is maximized when $p_c = 0.5$ and minimized when $p_c = 1.0$ or 0.0.

forecast and observations were uncorrelated. On island 2, the observed and forecast temperatures were drawn from *N*(0,β) distributions, and forecast and observed were correlated at 0.9. β varied between 1 and 3. Other aspects such as the ensemble size and number of days were the same as in the previous experiment.

As β increased, at island 2 the forecast and observed event frequency increased (Fig. 2). Ideally, the reported composite skills using the conventional methods would not change much as β changed, for the forecast-observed correlation never changed even though island 2's spread changed.

Figures 3 a-c shows that on island 1, skills remained near zero in each of the three metrics when using the conventional methods. On island 2, skill was near 1.0, and increased (*BSS* and *ETS*) or decreased (*ROCSS*) slightly with increasing β. When combined over both islands, the overall skill increased as β increased, as did the climatological event frequency and the uncertainty. Hence, the conventional methods apparently more heavily weighted the contribution from island 2 as β increased.

An examination of contingency tables for deterministic forecasts illuminates why the overall *ETS* was more heavily weighted toward island 2's contribution (Tables 2 – 4). Table 2 reports island 1's contingency table, Table 3 reports island 2's when β = 1, and Table 4 reports island 2's when β = 3. The *ETS* for island 1 alone was -0.0022, the *ETS* for island 2 and β = 1 is 0.4195, and the combined *ETS* when β = 1 was 0.193, nearly equally weighting the two islands' contributions. Note that when β = 1 the climatological event frequency was 0.0232 at island 1 and 0.0288 at island 2, very similar. However, for β = 3, the climatological event frequency at island 2 was 0.26, and its *ETS* was 0.532.

The combined *ETS* for β = 3 was 0.499, much closer to that of island 2 than 1. The unequal weighting is illuminated by considering the sums of the contingency tables. Note for example that the "hits" in the combined table for β = 3 (combining Tables 2 and 4) were determined almost exclusively by the hits from island 2, which contributed more than 98 percent.

This second example showed another undesirable property of the conventional method of calculating verification scores, namely that the weighting of samples is related to the observed event uncertainty. This may distort the calculation skill of important variables like heavy precipitation (see example in the appendix of Hamill and Whitaker 2006). In locations where heavy precipitation is quite rare (observational uncertainty small), the climatological reference produces a low-error forecast in most circumstances, and so a modest absolute forecast error can be evaluated as having negative skill relative to the climatology. Conversely, if heavy precipitation is more common (observational uncertainty larger), that same modest absolute forecast error may translate to a forecast with skill relative to the climatology, which is longer producing a low-error forecast in most circumstances. Hence the locations diagnosed as having more skill commonly ones with greater observational uncertainty; consequently, they may end up being more highly weighted in the calculation of the skill score, resulting in a skill larger than the average of skills at the constituent grid points.

It is also conceptually possible that skill could be underestimated using the conventional methods. This would have happened, for example, if this experiment was repeated, but this time forecasts and observations were highly correlated at Island 1 rather than Island 2. Practically, though, our experience suggests that skill tends to be more commonly overestimated (ibid).

The solutions proposed in the previous example may be useful here as well, with one exception. In this example the calculation of the *BSS* cannot be fixed by defining BS_c using eq. (7); it will yield a similar result to when eq. (5) is used. Equation (7) will still effectively weight the samples with greater climatological uncertainty higher than samples with less climatological uncertainty. If eq. (9) were used, the reported *BSS* would be a simple arithmetic average of the skill at the two islands. Similarly, the reported *ROCSS* will be an arithmetic average if calculated with eqs. (13) – (16), as will be the *ETS* if calculated separately at each island and then averaged using eq. (19).

4. **Example of skill overestimation: equitable threat scores for numerical precipitation forecasts**

Here we demonstrate that the *ETS* for real precipitation forecasts is subject to the same overestimation problem as with the synthetic data. The *ETS* is commonly used by the US National Weather Service to evaluate the skill of their deterministic precipitation forecasts. Typically, the ETS is estimated at fixed precipitation thresholds from a single contingency table populated over many days or months and over a wide geographic region such as the conterminous USA.

To demonstrate the tendency to report a larger-than-expected *ETS*, a very large set of numerical forecasts was used. These forecasts were generated using the analog forecast technique discussed in Hamill et al. (2006). The details of the forecast methodology can be found in this reference but are not particularly important here. What is relevant is that a 25-year time series of gridded deterministic precipitation forecasts was produced, all using the same model and forecast technique. These forecasts have characteristics roughly similar to those of current operational forecasts. For this demonstration, we limit ourselves to considering the *ETS* of the mean of a 5-member ensemble of analog forecasts over the conterminous USA (CONUS) for January and February from 1979 to 2003. Both the forecast and the verification data (from the North American Regional Reanalysis, Mesinger et al. 2005) are on a ~32 km grid. We consider the 5-mm precipitation threshold.

Figure 4a illustrates the geographic dependence of the *ETS* on forecast location. Contingency tables and *ETS* were calculated separately for each grid point. The *ETSs* were much larger in the southeast USA and along the West Coast than in the northwestern Great Plains. Figure 4b provides the climatological event frequency of greater than 5 mm rain in the 24-h period. Note the strong relationship between the *ETS* and the event frequency, a characteristic previously described for a similar skill score in Mason (1989) and for the *ETS* in Göber et al. (2004). Since observational uncertainty is thus typically larger at grid points with higher *ETS*, we might expect to see the effect demonstrated in section 3b, whereby an *ETS* calculated from the sum of all contingency tables across the CONUS will unduly weight the influence of the forecasts with the higher skill. Indeed, the *ETS* calculated from the contingency table sum using eq. (17) was approximately 0.415. However, examining Fig. 4a, it was apparent that the large majority of grid points had *ETS* much below 0.415. When calculated using eq. (19) after binning the climate into 6 categories[3], the weighted-average *ETS* was much smaller, ~ 0.28 (Fig. 5).

The *ETS* estimation technique of eq. (19) has drawbacks. Notably, the climatological event probability was defined by the sample event probability $(a+c)/(a+b+c+d)$, a reasonable assumption in this example with over two decades of winter forecast data, a very large sample. If the verification period is very short, then this sample event probability may be a poor estimate of the true long-term event probability. Ideally a long, temporally and spatially dependent climatology should be used, if available. If this were not possible, cross-validation techniques could be used to isolate the data being verified from the data being used to define the climatological event frequency. Nonetheless, these details should not obscure the main point: a substantially larger-than-expected threat score is possible when contingency table values are summed across grid points with different climatologies.

5. **Conclusions**

The preceding examples have demonstrated that the Brier skill score, relative operating characteristic, and the equitable threat score must be interpreted with care when verifying weather forecasts. These metrics, when conventionally applied, may report different skill than one would expect in situations where the climatological event frequency differs between sample locations. The more the event frequencies differ, the more the skill may be misestimated. By logical extension, skill may also be misestimated if the verification samples span different seasons or even different times of the day with

[3] Further subdivision into a greater number of categories did not increase the *ETS* appreciably.

different climatologies but the data are still composited. Other scores such as the ranked probability skill score and other contingency-table based scores can be assumed to be subject to the same tendencies. These misestimates can complicate the evaluation of model performance. Are two models nearly equal in their large skill because they're both providing high-quality forecasts? Or are they actually less skillful, and are differences in skill obscured by fictitious added skill from the varying climatology?

One primary reason why skill scores have been calculated as sums over sets with varying climatologies is that the sample size of the forecasts and observations are often small, and skills for the subsets may have a large sampling variability. A weighted-average skill over these subsets, as we have proposed, may not be resistant to outliers. Also, if independent observational data are not available to define the climatological event frequency for sub-samples, then this must be estimated from the same data used for model verification, potentially causing several additional problems. First, the small sample size may result in large errors in estimating the climatological event frequency. Second, unless the observational data used to define the climatology are separated from the observational data used for forecast verification to preserve independence (cross validation), the forecast skill may be underestimated; the error of the climatology will be diagnosed as smaller since it increasingly resembles the observed data as sample size decreases.

Clearly, talents of the statistical meteorologist will be put to the test when data are limited. While each situation may be different, one consideration should at least be to design the verification method to minimize the reported increase in skill introduced by varying climatologies, making at least relative inferences of skill (is model A more skillful than model B?) more trustworthy.

We propose two changes that both address the tendency for misestimating skill. First, if sample sizes are large enough, perform the calculations separately each for sub-sample with similar climatological event frequencies, as demonstrated for the equitable threat score in section 4. If the statistical meteorologist requires a single-number summary of the skill, consider weighted-average calculations similar to those proposed in section 2. Second, consider estimating skills for alternative events where the climatological event frequencies are the same for all samples, such as the exceedance of a quantile of the local climatological distribution (e.g., Buizza et al. 2003, Fig. 5, or Zhu et al. 2002). Then regardless of whether the climatological means and variances are large or small, the fraction events classified as "yes" events then are identical for different locations or times of the year. These recommendations are generally consistent with the recommendations published by the World Weather Research Program (WWRP) Working Group on Numerical Experimentation (WGNE) Joint Working Group on Verification (see http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/jwgv/jwgv.html).

A final recommendation is that the *specific details* regarding how the verification metrics are calculated should be fully described in journal articles and texts, since minor changes in the methodology can dramatically change the reported scores. Finally, whatever the chosen verification metric, it is prudent to verify that climatological forecasts report the expected no-skill result before proceeding.

Acknowledgments

Three anonymous reviewers are thanked for their constructive criticism. William “Matt” Briggs (Cornell University) is thanked for his review and for pointing out the Mantel-Haenszel test. Dan Wilks (Cornell), Craig Bishop (US Navy/NRL), Beth Ebert (Australian BOM), Steve Mullen (U. Arizona), Simon Mason (IRI), Bob Glahn (NOAA/MDL), Neill Bowler and Ken Mylne (UK Met Office), Bill Gallus (Iowa State), Frederic Atger (Meteo France), Francois LaLaurette (EMCWF and Meteo France), Zoltan Toth (NCEP), and Jeff Whitaker (NOAA/ESRL/PSD) are thanked for their discussions and comments on an earlier version of this manuscript. This research was supported by National Science Foundation grants ATM-0130154 and ATM-0205612.

References

Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2003: Sensitivity of precipitation forecast skill scores to bilinear interpolation and simple nearest-neighbor average method on high-resolution verification grids. *Weather and Forecasting*, **18**, 918-932.

Agresti, A., 2002: *Categorical Data Analysis*. Wiley Interscience, 710 pp.

Atger, F., 2003: Spatial and interannual variability of reliability of ensemble-based probabilistic forecasts: consequences for calibration. *Mon. Weather Rev.,* **131**, 1509-1523.

Bayler, G. M., R. M. Aune, and W. H. Raymond, 2000: NWP cloud initialization using GOES sounder data and improved modeling of nonprecipitating clouds. *Mon. Weather Rev.,* **128**, 3911–3920.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probabilities. *Mon. Weather Rev.,* **78**, 1-3.

Bright, D. R., and S. L. Mullen, 2002: Short-range ensemble forecasts of precipitation during the southwest monsoon. *Weather and Forecasting*, **17**, 1080–1100.

Buizza, R., and T. N. Palmer. 1998: Impact of ensemble size on ensemble prediction. *Mon. Weather Rev.,* **126**, 2503–2518.

----------, T. Petroliagis, T. N. Palmer, J. Barkmeijer, M. Hamrud, A. Hollingsworth, A. Simmons, and N. Wedi, 1998: Impact of model resolution and ensemble size on the performance of an ensemble prediction system. *Q. J. R. Meteorol. Soc.,* **124**, 1935-1960.

----------, A. Hollingsworth, F. Lalaurette, and A. Ghelli, 1999: Probabilistic predictions of precipitation using the ECMWF ensemble prediction system. *Wea. Forecasting*, **14**, 168-189.

----------, --------- , ---------, and --------, 2000a: Reply to comments by Wilson and by Juras. *Weather and Forecasting,* **15**, 367-369.

----------, J. Barkmeijer, T. N. Palmer, and D. S. Richardson, 2000b: Current status and future development of the ECMWF ensemble prediction system. *Meteorol. Appl.*, **7**, 163-175.

----------, 2001: Accuracy and potential economic value of categorical and probabilistic forecasts of discrete events. *Mon. Weather Rev.,* **129**, 2329-2345.

----------, D. S. Richardson, and T. N. Palmer, 2003: Benefits of increased resolution in the ECMWF ensemble prediction system and comparison with poor-man's ensembles. *Q. J. R. Meteorol. Soc.,* **129**, 1269-1288.

Chien, F.-C., Y.-H. Kuo, and M,-J. Yang, 2002: Precipitation forecast of MM5 in the Taiwan area during the 1998 Mei-yu season. *Weather and Forecasting*, **17**, 739–754.

Cochran, W. G., 1954: Some methods of strengthening the common χ^2 tests. *Biometrics*, **10**, 417-451.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Weather Rev.,* **129**, 2461–2480.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteorol.*, **8**, 190-198.

Gallus, W. A., Jr., and M. Segal, 2001: Impact of improved initialization of mesoscale features on convective system rainfall in 10-km Eta simulations. *Wea. Forecasting*, **16**, 680–696.

-----------, and ----------, 2004: Does increased predicted warm-season rainfall indicate enhanced likelihood of rain occurrence? *Weather and Forecasting*, **19**, 1127–1135.

Gandin, L. S., and A. H. Murphy, 1992: Equitable skill scores for categorical forecasts. *Mon. Weather Rev.,* **120**, 361-370.

Glahn, B., 2004: Discussion of verification concepts in "Forecast Verification: A Practitioner's Guide in Atmospheric Science." *Weather and Forecasting*, **19**, 769-775.

Göber, M., C. A. Wilson, S. F. Milton, and D. B. Stephenson, 2004: Fairplay in the verification of operational quantitative precipitation forecasts. *J. Hydrol.*, **288**, 225-236.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Weather and Forecasting*, **14**, 155-167.

----------- , C. Snyder , and R. E. Morss, 2000: A comparison of probabilistic forecast from bred, singular vector and perturbed observation ensembles. *Mon. Weather Rev.,* **128**, 1835-1851.

----------- , J. S. Whitaker, and S. L. Mullen, 2006: Reforecasts, an important new data set for improving weather predictions. *Bull. Amer. Meteorol. Soc.*, **87**, 33-46.

----------- , and ----------- , 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: theory and application. *Mon. Wea. Rev*., in press. Available at www.cdc.noaa.gov/people/tom.hamill/reforecast_analog_v2.pdf

Harvey, L. O., Jr., K. R. Hammond, C. M. Lusk, and E. F. Mross, 1992: The application of signal detection theory to weather forecasting behavior. *Mon. Weather Rev.,* **120**, 863-883.

Juras, J., 2000: Comments on "probabilistic predictions of precipitation using the ECMWF ensemble prediction system." *Weather and Forecasting,* **15**, 365-366.

Kharin, V. V., and F. W. Zwiers, 2003: On the ROC score of probability forecasts. *J. Climate*, **16**, 4145–4150.

Kheshgi, H. S., and B. S. White, 2001: Testing distributed parameter hypotheses for the detection of climate change. *J. Climate*, **14**, 3464–3481.

Legg, T. P., K. R. Mylne. 2004: Early warnings of severe weather from ensemble forecast information. *Weather and Forecasting*, **19**, 891–906.

Mantel, N. and W. Haenszel, 1959: Statistical aspects of the analysis of data from retrospective studies of disease. *J. National Cancer Institute*, **22**, 719-748.

Marzban, C. 2004: The ROC curve and its area under it as performance measures. *Weather and Forecasting*, **19**, 1106-1114.

Mason, I. B., 1982: A model for the assessment of weather forecasts. *Aust. Meteorol. Mag*., **30**, 291-303.

---------- , 1989: Dependence of the critical success index on sample climate and threshold probability. *Aust. Meteorol. Mag*., **37**, 75-81.

---------- , 2003: Binary events. In *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* I. T. Jolliffe and D. B. Stephenson, Eds., John Wiley, 37-76.

Mason, S. J., and N. E. Graham, 1999: Conditional probabilities, relative operating characteristics, and relative operating levels. *Weather and Forecasting*, **14**, 713-725.

-----------, and -----------, 2002: Areas beneath relative operating characteristics (ROC) and relative operating levels (ROL) curves: statistical significance and interpretation. *Q. J. R. Meteorol. Soc.,* **128**, 2145-2166.

----------- , 2004: On using "climatology" as a reference strategy in the Brier and ranked probability skill scores. *Mon. Weather Rev.,* **132**, 1891-1895.

Mesinger, F., and coauthors, 2005: North American regional reanalysis. *Bull. Amer. Meteorol. Soc*., submitted. Available from fedor.mesinger@noaa.gov .

Mullen, S. L., and R. Buizza, 2001: Quantitative precipitation forecasts over the United States by the ECMWF ensemble prediction system. *Mon. Weather Rev.,* **129**, 638–663.

-----------, and ----------, 2002: The impact of horizontal resolution and ensemble size on probabilistic forecasts of precipitation by the ECMWF ensemble prediction system. *Weather and Forecasting*, **17**, 173–191.

Murphy, A. H., 1971: A note on the ranked probability score. *J. Appl. Meteorol.*, **10**, 155-156.

Palmer, T. N., C. Brankovic, and D. S. Richardson, 2000: A probability and decision-model analysis of PROVOST seasonal multi-model ensemble integrations. *Q. J. R. Meteorol. Soc.,* **126**, 2013-2033.

Richardson, D. S. , 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.,* **126**, 649-667.

----------, 2001a: Ensembles using multiple models and analyses. *Q. J. R. Meteorol. Soc.,* **127**, 1847-1864.

----------, 2001b: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Q. J. R. Meteorol. Soc.,* **127**, 2473-2489.

Rogers, E., D. G. Deaven, and G. J. DiMego, 1995: The regional analysis system for the operational "early" Eta Model: original 80-km configuration and recent changes. *Wea. Forecasting*, **10**, 810-825.

-----------, and coauthors, 1996: Changes to the operational "early" Eta analysis/forecast system at the National Centers for Environmental Prediction. *Wea. Forecasting*, **11**, 391-413.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Weather and Forecasting*, **5**, 570-575.

Simpson, E. H., 1951: The interpretation of interaction in contingency tables. *J. Royal. Stat. Soc.*, **13**, 238-241.

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989. *Survey of common verification methods in meteorology*. Enviroment Canada Research Report 89-5, 114 pp.

Available from Atmospheric Environment Service, Forecast Research Division, 4905 Dufferin St., Downsview, ON M3H 5T4, Canada.

Stefanova, L., and T. N. Krishnamurti, 2002: Interpretation of seasonal climate forecasts using Brier skill score, the Florida State University superensemble and the AMIP-I data set. *J. Climate*, **15**, 537-544.

Stensrud, D. J., J.-W. Bao, and T. T. Warner, 2000: Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Weather Rev.,* **128**, 2077-2107.

Swets, J. A., 1973: The relative operating characteristic in psychology. *Science*, **182**, 990-999.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. Chapter 7 of "*Forecast Verification: A Practitioner's Guide in Atmospheric Science*." John Wiley and Sons, 254 pp.

Wandishin, M. S., S. L. Mullen, D. J. Stensrud, and H. E. Brooks. 2001: Evaluation of a short-range multimodel ensemble system. *Mon. Weather Rev.,* **129**, 729–747.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences (2nd Ed)*. Academic Press. 627 pp.

-----------, 2001: A skill score based on economic value for probability forecasts. *Meteorol. Appl.,* **8**, 209-219.

Wilson, L. J., 2000: Comments on "probabilistic predictions of precipitation using the ECMWF ensemble prediction system." *Weather and Forecasting,* **15**, 361-364.

World Meteorological Organization, 1992: *Manual on the Global Data Processing System*, section III, Attachment II.7 and II.8, (revised in 2002). Available from http://www.wmo.int/web/www/DPS/Manual/WMO485.pdf.

Xu, M., D. J. Stensrud, J.-W. Bao, and T. T. Warner, 2001: Applications of the adjoint technique to short-range forecasting of mesoscale convective systems. *Mon. Weather Rev.,* **129**, 1395-1418.

Yang, Z., and R.W. Arritt, 2002: Tests of a perturbed physics ensemble approach for regional climate modeling. *J. Climate*, **15**, 2881–2896.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson and K. R. Mylne. 2002: The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteorol. Soc*., **83**, 73–83.

LIST OF FIGURES

Figure 1: *ROCSS, BSS*, and *ETS* as a function of the parameter α, which describes the difference in the means of the distributions between the two islands.

Figure 2: Illustration of experimental design (section 3b). Weather forecasts are simulated at two islands, the first island with uncorrelated forecasts and observations, the second island with forecasts and observations correlated at 0.90. The observed and forecast data are normally distributed with zero mean. The standard deviation β is fixed at 1.0 for island 1 and varies between 1 and 3 at island 2. (a) $\beta = 1.0$, (b) $\beta = 3.0$. Dotted lines indicates the event threshold (observed or forecast temperature greater than 2.0). The three contours enclose 90, 50, and 10 percent of the probability density, respectively.

Figure 3: (a) *BSS*, (b) *ETS*, and (c) *ROC AUC* at individual islands and when combined using the conventional methods for experiment in section 3(b).

Figure 4: (a): *ETS* for 1-2 day (24-48 h) 5 mm precipitation forecasts as a function of location, using Jan-Feb 1979-2003 forecast and analyzed data. (b) Climatological probability of precipitation greater than 5 mm for Jan-Feb.

Figure 5: Histogram of *ETS* of 5-mm forecasts when for subsets of samples divided into six categories based on the climatological probability of event occurrence. The fraction of the grid points occurring in a given bin are reported in parentheses. Dashed lines

indicate the *ETS* calculated using the conventional method (eq. (17)) and the population-weighted average (eq (19)).

LIST OF TABLES

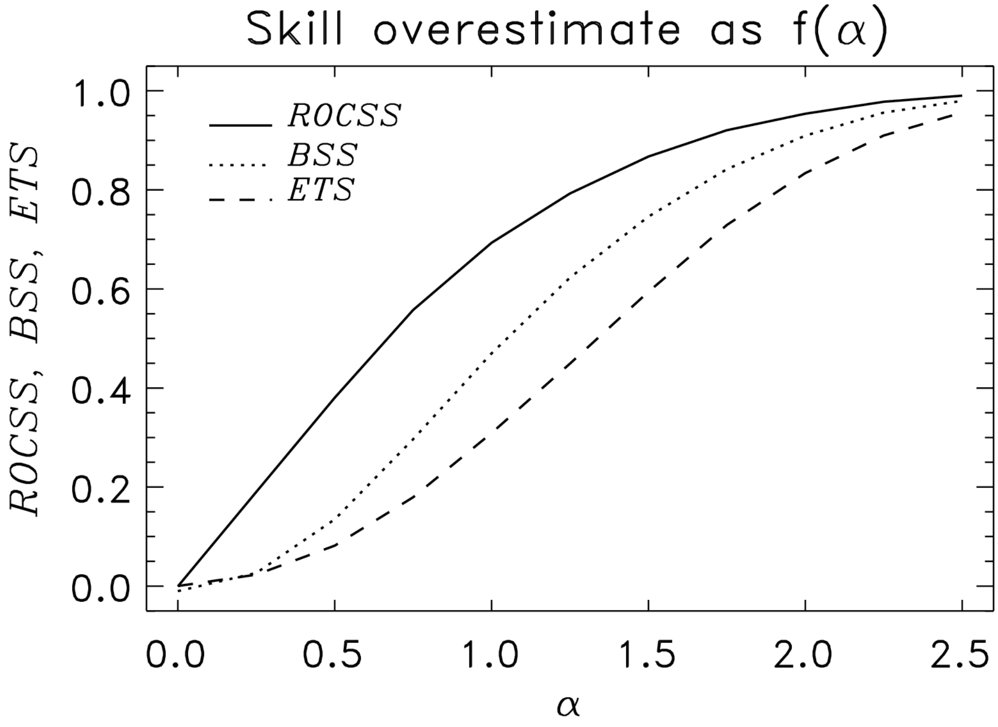


Figure 1: *ROCSS, BSS*, and *ETS* as a function of the parameter α, which describes the difference in the means of the distributions between the two islands.

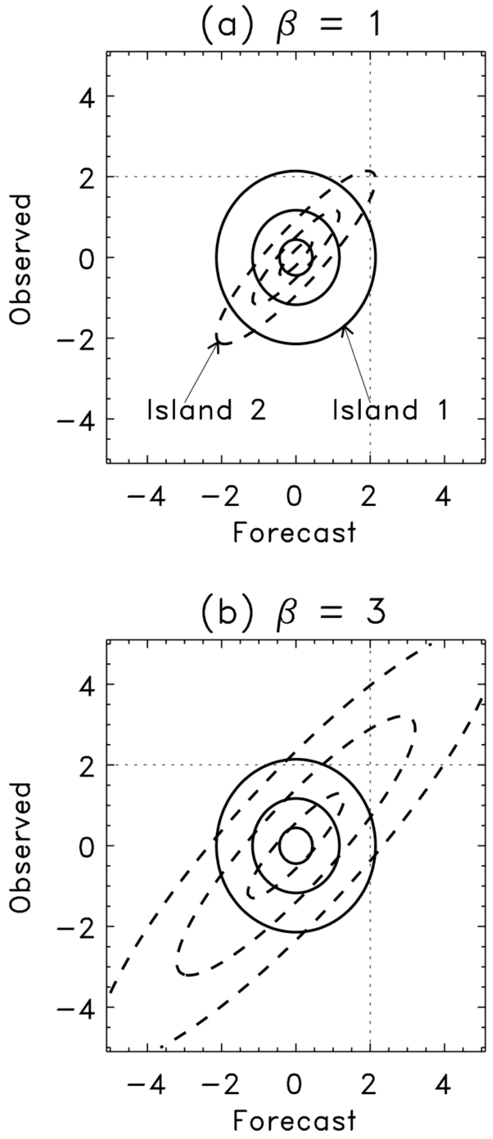


Figure 2: Illustration of experimental design (section 3b). Weather forecasts are simulated at two islands, the first island with uncorrelated forecasts and observations, the second island with forecasts and observations correlated at 0.90. The observed and forecast data are normally distributed with zero mean. The standard deviation β is fixed at 1.0 for island 1 and varies between 1 and 3 at island 2. (a) β = 1.0, (b) β = 3.0. Dotted line indicates the event threshold (observed or forecast temperature greater than 2.0). The three contours enclose 90, 50, and 10 percent of the probability density, respectively.

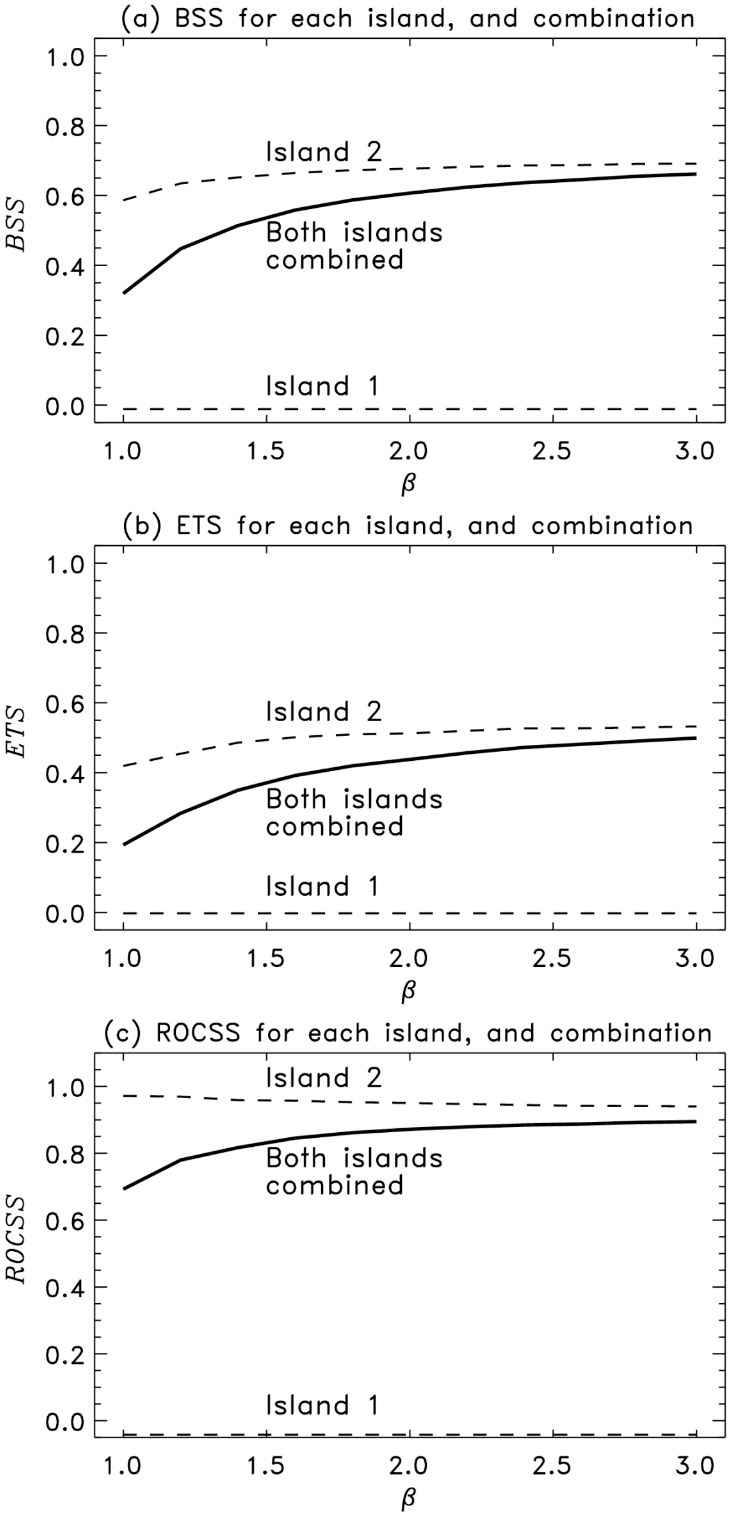
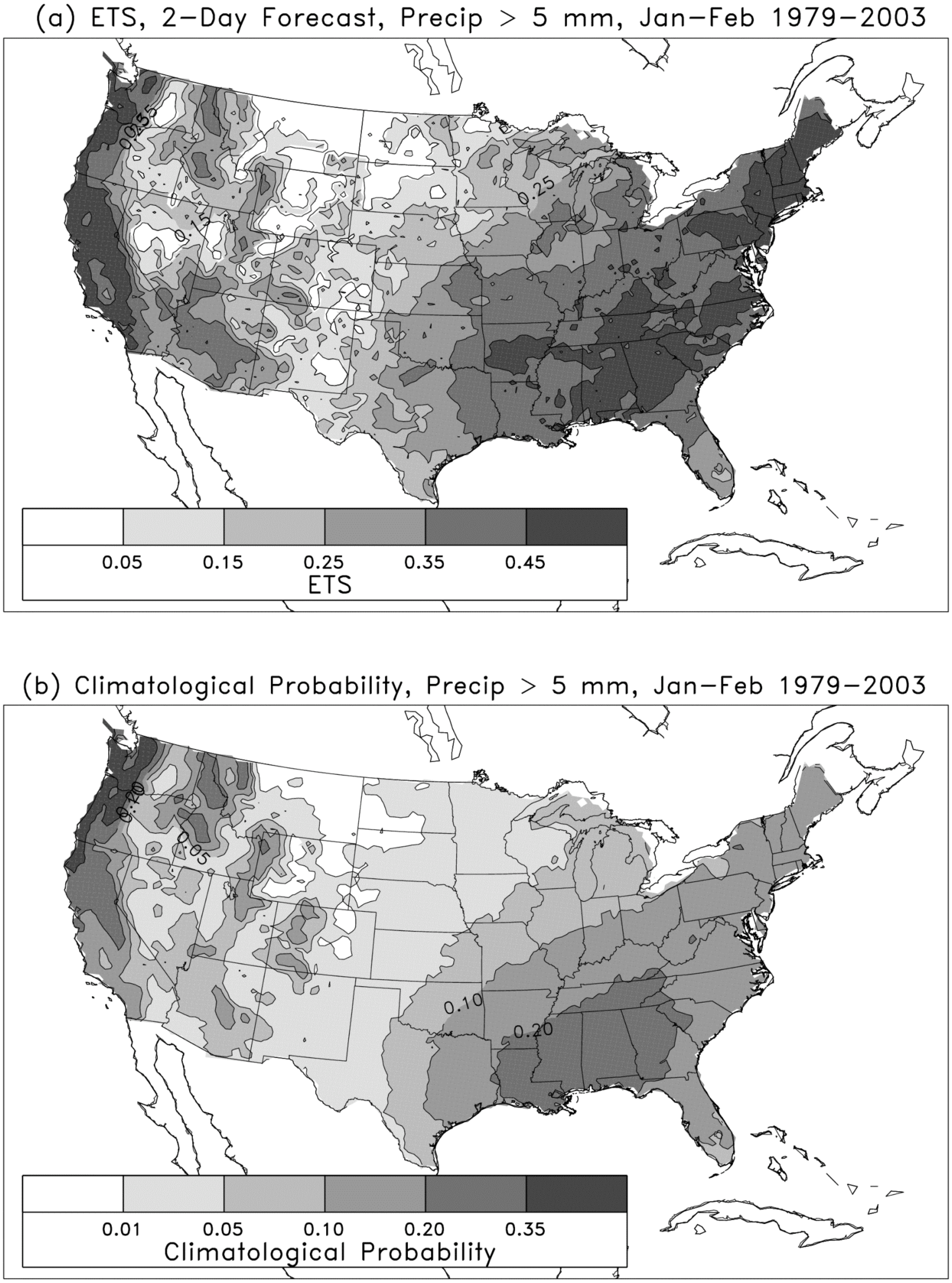


Figure 3: (a) *BSS*, (b) *ETS*, and (c) *ROCSS* at individual islands and when combined using the conventional methods for experiment in section 3(b).



(a) ETS, 2–Day Forecast, Precip > 5 mm, Jan–Feb 1979–2003

(b) Climatological Probability, Precip > 5 mm, Jan–Feb 1979–2003

Figure 4: (a): *ETS* for 1-2 day (24-48 h) 5 mm precipitation forecasts as a function of location, using Jan-Feb 1979-2003 forecast and analyzed data. (b) Climatological probability of precipitation greater than 5 mm for Jan-Feb.

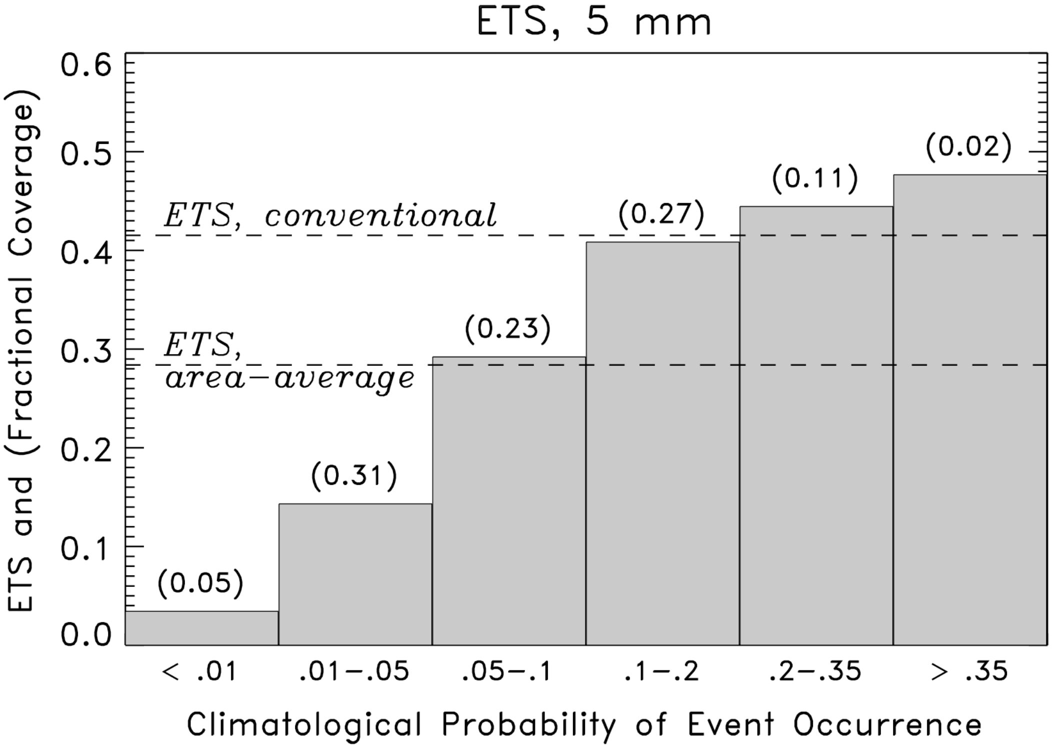


Figure 5: Histogram of *ETS* of 5-mm forecasts when for subsets of samples divided into six categories based on the climatological probability of event occurrence. The fraction of the grid points occurring in a given bin are reported in parentheses. Dashed lines indicate the *ETS* calculated using the conventional method (eq. (17)) and the population-weighted average (eq (19)).

		Event Observed?	
		YES	NO
Event forecast by *i*th member?	YES	a_i	b_i
	NO	c_i	d_i

Table 1: Contingency table for the *i*th of the *n* sorted members at the *j*th location, indicating the relative fraction of hits [a_i], false alarms [b_i], misses [c_i], and correct rejections [d_i].

		Event Observed?	
		YES	NO
Event Forecast?	YES	0.004	0.0223
	NO	0.0228	0.954

Table 2: Contingency table for island 1 in experiment in section 3 (b). The observed event frequency is 0.0232 and the *ETS* is – 0.0022.

		Event Observed?	
		YES	NO
Event Forecast?	YES	0.0171	0.0108
	NO	0.0117	0.9603

Table 3: Contingency table for island 2 in experiment in section 3 (b) when $\beta = 1.0$. The observed event frequency is 0.0288 and the *ETS* is +0.4195.

		Event Observed?	
		YES	NO
Event Forecast?	YES	0.2022	0.0597
	NO	0.0578	0.6802

Table 4: Contingency table for island 2 in experiment in section 3 (b) when $\beta = 3.0$. The observed event frequency is 0.26 and the *ETS* is +0.5327.